# Discovery of biomarkers for glycaemic deterioration before and after the onset of type 2 diabetes: descriptive characteristics of the epidemiological studies within the IMI DIRECT Consortium

Robert W. Koivula[1,2], Ian M. Forgie[3], Azra Kurbasic[1], Ana Viñuela[4,5,6], Alison Heggie[7], Giuseppe N. Giordano[1], Tue H. Hansen[8], Michelle Hudson[9], Anitra Koopman[10], Femke Rutters[10], Maritta Siloaho[11], Kristine H. Allin[8,12], Søren Brage[13,14], Caroline A. Brorsson[15], Adem Y. Dawed[3], Federico De Masi[15], Christopher J. Groves[2], Tarja Kokkola[11], Anubha Mahajan[16], Mandy H. Perry[9], Simone P. Rauh[10], Martin Ridderstråle[17,18], Harriet J. A. Teare[19], E. Louise Thomas[20], Andrea Tura[21], Henrik Vestergaard[8], Tom White[13], Jerzy Adamski[22], Jimmy Bell[20], Joline Beulens[10], Søren Brunak[15,23], Emmanouil T. Dermitzakis[4,5,6], Philippe Froguel[24,25], Gary Frost[26], Ramneek Gupta[15], Torben Hansen[8,14], Andrew Hattersley[9,27], Bernd Jablonka[28], Jayne Kaye[19], Markku Laakso[11], Timothy J. McDonald[9], Oluf Pedersen[8], Jochen M. Schwenk[29], Imre Pavo[30], Andrea Mari[21], Mark I. McCarthy[2,16,31], Hartmut Ruetten[28], Mark Walker[7], Ewan Pearson[3], Paul W. Franks[1,2,32,33], *for the IMI DIRECT Consortium*

**Affiliations:**

1. Department of Clinical Sciences, Genetic and Molecular Epidemiology Unit, Skåne University Hospital Malmö, Lund University, Malmö, Sweden.
2. Oxford Centre for Diabetes, Endocrinology and Metabolism, Radcliffe Department of Medicine, University of Oxford, Oxford, United Kingdom.
3. Population Health & Genomics, School of Medicine, University of Dundee, Ninewells Hospital, Dundee, United Kingdom.
4. Department of Genetic Medicine and Development, University of Geneva Medical School, Geneva, Switzerland.

5. Institute of Genetics and Genomics in Geneva (iGE3), University of Geneva, Geneva, Switzerland.
6. Swiss Institute of Bioinformatics, Geneva, Switzerland.
7. Institute of Cellular Medicine (Diabetes), Newcastle University, Newcastle upon Tyne, United Kingdom.
8. The Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health and Medical Science, University of Copenhagen.
9. NIHR Exeter Clinical Research Facility, University of Exeter Medical School, Exeter, UK.
10. Department of Epidemiology and Biostatistics, Amsterdam Public Health Research Institute, VU University Medical Center, Amsterdam, the Netherlands.
11. Department of Medicine, University of Eastern Finland and Kuopio University Hospital, Finland.
12. Department of Clinical Epidemiology, Bispebjerg and Frederiksberg Hospital, the Capital Region, Copenhagen, Denmark.
13. MRC Epidemiology Unit, University of Cambridge School of Clinical Medicine, Cambridge, UK.
14. Faculty of Health Sciences, University of Southern Denmark, Odense, Denmark.
15. Department of Bio and Health Informatics, Technical University of Denmark, Lyngby, Denmark.
16. Wellcome Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK.
17. Department of Clinical Sciences, Clinical Obesity, Skåne University Hospital Malmö, Lund University, Malmö, Sweden.
18. Novo Nordisk A/S, Søborg, Denmark.

19. HeLEX, Nuffield Department of Population Health, University of Oxford, Old Road Campus, Headington, Oxford, UK.
20. Research Centre for Optimal Health, Department of Life Sciences, University of Westminster, London, U.K.
21. Institute of Neurosciences, National Research Council, Padova, Italy.
22. Institute of Epidemiology II, Helmholtz Zentrum Muenchen, German Research Center for Environmental Health (GmbH), Neuherberg, Germany.
23. The Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, Copenhagen, Denmark.
24. Department of Genomics of Common Disease, School of Public Health, Imperial College London, London, United Kingdom.
25. CNRS, Pasteur Institute of Lille, University of Lille, Lille, France.
26. Nutrition and Dietetics Research Group, Department of Medicine, Division of Diabetes, Endocrinology and Metabolism, Imperial College London, Hammersmith Campus, London W12 0NN, UK.
27. Institute of Biomedical and Clinical Science, University of Exeter Medical School, Exeter, UK.
28. Sanofi-Aventis Deutschland GmbH, R&D, Frankfurt am Main, Germany.
29. Science for Life Laboratory, School of Engineering Sciences in Chemistry, Biotechnology and Health, KTH - Royal Institute of Technology, Stockholm, Sweden.
30. Eli Lilly Regional Operations GmbH, Vienna, Austria.
31. NIHR Oxford Biomedical Research Centre, Churchill Hospital, Oxford, UK.
32. Department of Nutrition, Harvard School of Public Health, Boston, MA, USA.

33. Department of Public Health & Clinical Medicine, Section for Medicine, Umeå University, Umeå, Sweden.

**Corresponding authors:**

Paul Franks:

Department of Clinical Sciences, Lund University, Genetic and

Molecular Epidemiology, CRC, Skåne University Hospital Malmö,

Building 91, Level 10, Jan Waldenströms gata 35,

SE-205 02 Malmö, Sweden

e-mail: paul.franks@med.lu.se

Ewan Pearson:

Population Health & Genomics, Medical Research Institute,

University of Dundee, Dundee DD1 9SY, UK

e-mail: e.z.pearson@dundee.ac.uk

**Abbreviations:**

ASAT: Abdominal subcutaneous adipose tissue

DIRECT: Diabetes Research on Patient Stratification

EU: European Union

MMTT: Mixed-meal tolerance test

MRI: Magnetic resonance imaging

hpfVM: High-pass filtered vector magnitude

IAAT: Intra-abdominal adipose tissue

IGR: Impaired glucose regulation

IMI: Innovative Medicines Initiative

ME: multiecho

NGR: Normal glucose regulation

OGTT: Oral glucose tolerance test

PA: Physical activity

TAAT: Total abdominal adipose tissue

**Key words:**

Prediabetes; Type 2 diabetes; Glycaemic control; Insulin secretion; Insulin sensitivity; Ectopic fat; Diet; Physical Activity; Genome; Personalized Medicine.

**Running Title:** IMI DIRECT epidemiological studies data overview

**Word Count:** Main body: 4552/4000, Abstract: 297, Research in Context: 194

**Figures and Tables:** 6 (+ 1 ESM table)

**Abstract**

*Objectives:* Here, we describe the characteristics of the IMI DIRECT (Diabetes Research on Patient Stratification) epidemiological cohorts at baseline and follow-up examinations (18, 36 and 48 month follow-up).

*Research design and methods:* From a sampling-frame of 24,682 European-ancestry adults enrolled in population-based cohorts across Europe, participants at varying risk of glycaemic deterioration were identified using a risk prediction algorithm (based on age, BMI, waist circumference, smoking status and family history of type 2 diabetes) and enrolled into a prospective cohort study (n=2127) (Cohort 1: prediabetes risk). We also recruited people from clinical registries with type 2 diabetes diagnosed 6 to 24 month prior (n=789) into a second cohort study (Cohort 2: diabetes). The cohorts were studied in parallel using matched protocols across seven clinical centres in northern Europe.

*Results:* Using ADA-2011 glycaemic categories, 33% (n=693) of Cohort 1 (prediabetes risk) had normal glucose regulation (NGR), and 67% (n=1419) had impaired glucose regulation (IGR). 76% of the cohort was male; age=62(6.2) years; BMI=27.9(4.0) kg/m$^2$; fasting glucose=5.7(0.6) mmol/l; 2-hr glucose=5.9(1.6) mmol/l [mean(SD)]. At follow-up, examinations took place ~18 and ~48 months after baseline; fasting glucose=6.0(0.6) mmol/l; 2-hr OGTT glucose=6.5 (2.0) mmol/l [mean(SD)] at final follow-up. In Cohort 2 (diabetes): 65% (n=508) were lifestyle treated (LS) and 35% (n=271) were lifestyle + metformin treated (LS+MET) at enrolment. 58% of the cohort was male; age=62(8.1) years; BMI=30.5(5.0) kg/m$^2$; fasting glucose=7.2(1.4)mmol/l; 2-hr glucose=8.6(2.8) mmol/l [mean(SD)]. At follow-up, examinations took place ~18 and ~36 months after baseline; fasting glucose=7.9(2.0) mmol/l; 2-hr MMTT glucose=9.9(3.4) mmol/l [mean(SD)] at final follow-up.

*Conclusions:* The IMI DIRECT cohorts are intensely characterized with a wide variety of metabolically relevant measures assessed prospectively. The cohorts aim to provide a powerful resource for biomarker discovery, multivariate aetiological analyses and reclassification of patients for the prevention and treatment of type 2 diabetes.

**Research in Context**

*What is already known about this subject?*

- Type 2 diabetes is a heterogeneous disease composed of several pathophysiological insults with the combined result of a loss of blood glucose control.
- Measures of these pathophysiological insults may facilitate the stratification of type 2 diabetes into treatable subclasses.

*What is the key question?*

- What biomarkers can be used to stratify and individuals at risk of- and with type 2 diabetes into separate subclasses of glycaemic deterioration?

*What are the new findings?*

- Cohort 1 (Prediabetes risk): 2127 participants at risk of rapid glycaemic deterioration were recruited from a sampling-frame of 24,682 participants.
- Cohort 2 (Diabetes): 789 participants with recently diagnosed type 2 diabetes were recruited from local registries and primary care centres at 6 European study centres.
- Both cohorts are richly phenotyped. These include: metabolic biochemistry, measures of glycaemic control (OGTT/MMTT), region/organ-specific adiposity (MRI), physical activity/sleep (accelerometry), diet (self-reported and objectively assessed), blood omic assessments (genomic, transcriptomic, metabolomic, proteomic), and faecal microbiomics.

*How might this impact on clinical practice in the foreseeable future?*

- The richly phenotyped IMI DIRECT cohorts will facilitate the discovery of biomarkers for glycaemic control in individuals at risk of- or with type 2 diabetes.

## Introduction

The global prevalence of type 2 diabetes is burgeoning. There is no cure, nor are there treatments effective enough to halt the progression of the disease. The burden the disease conveys at a societal and personal level is enormous, with an estimated world prevalence in 2017 of around 425 million people with type 2 diabetes and a further 352 million at risk of developing the disease[1]. The global cost of diagnosing and treating the disease and its complications in 2017 was estimated to be around €730 billion[1]. This bleak picture emphasizes the profound shortcomings in our understanding of type 2 diabetes aetiology and pathogenesis, and the inadequate tools available with which to combat the disease.

Like some other complex diseases, the clinical presentation and prognosis of type 2 diabetes is heterogeneous. The risk conveyed by established diabetogenic factors such as obesity, physical inactivity and certain dietary components varies widely from one person to the next, as does their response to interventions targeting these risk factors. This is also true for those in whom diabetes is manifest, with response to glucose-lowering therapies, occurrence of adverse events and rates of progression being variable and hard to predict.

The diagnosis of type 2 diabetes is relatively straightforward, relying primarily on evidence of chronically elevated blood glucose concentrations [2]. However, elevated blood glucose concentrations can be the consequence of multiple defects in energy metabolism occurring across several organs and tissues [3-5] caused by myriad acquired or inherited factors. Thus, type 2 diabetes as currently defined, characterizes a collection of underlying pathologies [6], each with the common feature of elevated blood glucose, that may require tailored therapies. The stratification of type 2 diabetes into treatable subclasses would be possible if accessible biomarkers of the disease's underlying pathologies were known.

Although improving the management of type 2 diabetes through sub-classification may lead to more focused treatment, susceptibility to risk factors and response to treatments also vary. Thus, stratifying patient populations into subgroups defined using biomarkers quantifying susceptibility to risk factors and responsiveness to specific therapeutics would further enhance our ability to treat, and ideally prevent, the disease.

The IMI DIRECT consortium is a collaboration among investigators from some of Europe's leading academic institutions and pharmaceutical companies [7]. The overarching objective of IMI DIRECT is to discover and validate biomarkers of glycaemic deterioration before and after the onset of type 2 diabetes. To this end, we established two new multi-centre prospective cohort studies comprised of adults from northern Europe at risk of, or with, recently diagnosed type 2 diabetes. Within these cohorts, a comprehensive array of risk factors, intermediate phenotypes, and metabolic outcomes are repeatedly assessed using cutting-edge technologies. The richly phenotyped IMI DIRECT cohorts will facilitate the discovery of biomarkers for glycaemic control in individuals at risk of- or with type 2 diabetes.

Here we describe the characteristics of the two IMI DIRECT cohorts at baseline and at the 2 major follow-up visits up to 48 months later, to provide context for those subsequently analysing and reviewing studies based on these data. We also consider these results in the context of the implemented protocols and plans outlined at the beginning of the project, as described previously [7].

**Methods:**

The rationale and design of the epidemiological cohorts within IMI DIRECT are reported elsewhere [7]; here we provide data and information about key variables and methods respectively that were not described in the rationale and design paper published previously.

Approval for the study protocol was obtained from each of the regional research ethics review boards separately and all participants provided written informed consent at enrolment. The research conformed to the ethical principles for medical research involving human subjects outlined in the declaration of Helsinki.

*Recruitment, enrolment and eligibility:* The derivations of Cohort 1 and Cohort 2 are shown in Figure 1. The sampling-frame for Cohort 1 comprised four existing prospective cohort studies: *Metabolic Syndrome in Men* (METSIM) (Finland) [8]; *Relationship between Insulin Sensitivity and Cardiovascular disease* (RISC) [9], *Hoorn Meal Study* (HMS), and *New Hoorn Study* (NHS) [10] (Netherlands); *Health2010* [11], *Health2006* [12], *Danish Study of Functional Disorders* (DanFunD) [13], and *Gut, Grain and Greens* (GGG) [14] studies (Denmark); *Malmö Diet and Cancer* (MDC) study (Sweden) [15]. Participants for Cohort 2 were identified through general practice and other registries, as described previously [7].

After excluding participants who did not meet the inclusion criteria or whose data failed quality control, a total of 2127 and 789 participants at risk of or with type 2 diabetes were retained in Cohort 1 and Cohort 2, respectively. For Cohort 1, emphasis was placed on recruiting participants deemed at high-risk of type 2 diabetes by ADA-2011 $HbA_{1c}$ criteria, i.e. those with $HbA_{1c}$ concentrations between 40-48 mmol/mol (5.7-6.4%)[2, 7]. As anticipated during the design phase, the sampling-frame contained too few participants that fulfilled this criterion; thus, we proceeded to enrol participants with progressively lower $HbA_{1c}$ concentrations, who were also considered at highest risk of glycaemic deterioration according to the DIRECT-DETECT risk algorithm applied to the parent cohort sampling-frame [7, 16].

In Cohort 1, 1989 (93%) participants enrolled at baseline also attended the first major follow-up visit at 18.6(1.4) months [mean(SD)]; and 1729(81%) attended the final follow-up visit

30.8(1.3) months later. Similarly, in Cohort 2, 668 (85%) participants enrolled at baseline attended their first major follow-up visit at 18.2(0.6) months [mean(SD)] ; and 552(70%) attended the final follow-up visit 18.2 (1.0) months later.

*Glycaemic biochemistry assays:* Plasma glucose, insulin and C-peptide assays for Cohort 1 were carried out centrally at the University of Eastern Finland (Kuopio, Finland), where plasma glucose was analysed using the enzymatic glucose hexokinase method and photometric measurement on Konelab 20 XT Clinical Chemistry analyser (Thermo Fisher Scientific, Vantaa, Finland). In Cohort 2, plasma insulin and C-peptide were analysed using chemiluminometric immunoassay, (CLIA), Liaison Insulin and Liaison C-peptide **(**DiaSorin S.p.A, Saluggia, Italy). The instrument used was DiaSorin Liaison Analyser, DiaSorin Deutschland GmbH, Dietzenbach, Germany. Plasma glucose, insulin and C-peptide assessments for Cohort 2 were carried out centrally at the University of Exeter (Exeter, UK). Assessments of $HbA_{1c}$, blood lipids, ALT and AST for both cohorts were carried out centrally at the University of Exeter. Glucose was measured by the enzymatic colorimetric assay GOD-PAP using Roche MODULAR P analysers (F. Hoffmann-La Roche AG, Basel, Switzerland). Insulin was measured by electrochemiluminescence using Roche E170 Analysers (F. Hoffmann-La Roche AG, Basel, Switzerland). C-peptide concentrations in plasma and urine were measured by electrochemiluminescence using Roche E170 Analysers (F. Hoffmann-La Roche AG, Basel, Switzerland). $HbA_{1c}$ was measured by ion-exchange High Performance Liquid Chromatography using Tosoh G8 analysers (Tosoh Bioscience Inc, CA, USA). In each case biochemical assays were performed using validated standard methods. Reference samples were included in all procedures to control for inter-assay variation and laboratories were regularly participating in international external quality assessment schemes. In addition, a sub-set of samples were assayed for C-peptide Insulin and glucose on both sites to assess inter-laboratory variation.

*Blood lipid and liver enzyme biochemistry assays:* Triglyceride was measured by quantitative determination with glycerol blanking. HDL cholesterol was measured directly using PEG-modified enzymes and dextran sulphate. When cholesterol esterase and cholesterol oxidase enzymes are modified by PEG they show selective catalytic activities toward lipoprotein fractions, with reactivity increasing in the order: LDL < VLDL ≈ chylomicrons < HDL. Total cholesterol was measured by an enzymatic, colorimetric method. LDL cholesterol was calculated from the total cholesterol, HDL cholesterol and triglyceride concentrations using the Friedewald Formula: LDL = total cholesterol - HDL cholesterol - (triglyceride/2.2). ALT and AST were measured by UV absorbance without pyridoxal phosphate activation. ALT, AST, cholesterol, glucose, triglycerides, and HDL cholesterol were measured using a Roche MODULAR P analyser (Roche Diagnostics, Inc, IN, USA). Insulin and C-peptide were measured using a Roche E170 analyser (Roche Diagnostics, Inc, IN, USA).

*Blood GLP-1 assays:* Plasma concentrations of GLP-1 were determined by drawing blood samples collected at two different time points (0 and 60 min) during the 75g fsOGTT/MMTT (baseline samples only). P800 tubes (Becton Dickenson, UK) were used to provide immediate protection from intrinsic proteolysis. Quantitative determination of active GLP-1 was achieved using MSD GLP-1 active kit (Meso Scale Diagnostics LLC, MD, USA. Product code: K150JWC). Total GLP-1 was assayed using MSD GLP-1 total kit (Meso Scale Diagnostics LLC, MD, USA. Product code: K150JVC).

*Abdominal MRI analyses:* The volume of adipose tissue was measured in litres using magnetic resonance imaging (MRI), as described elsewhere [17]. *Total abdominal adipose tissue* (TAAT) may be separated into *intra-abdominal adipose tissue* (IAAT – also known as 'visceral' fat) and *abdominal subcutaneous adipose tissue* (ASAT). IAAT is the volume of adipose tissue within the abdominal cavity. TAAT is the sum of IAAT and ASAT. Liver and

pancreas fat and iron (T2*) were derived simultaneously using a multiecho (ME) MRI technique, as previously described [17, 18]. This method has the advantage over single voxel MRS in that regional differences in ectopic fat distribution can be measured. Furthermore, it is often possible to obtain a single slice quantification of the liver and pancreas, allowing simultaneous measurement of fat and iron within two separate organs. A biexponential curve-fitting model was used to derive the relative signal contributions from fat and water from the many images normally obtained with the ME sequence. Briefly, tissue with no fat infiltration generates a very smooth decay curve, whereas tissue containing a higher level of fat is characterized by significant oscillations throughout the decay curve [18]. A further output from the ME technique is T2* tissue values; as changes in these are indicative of iron content it provides a clinically relevant additional measurement. Tissue iron concentration (Fe, units mg/g dry weight tissue) was estimated from T2* using a validated model [19].

*Diet assessment:* Self-reported dietary intake was assessed by a 24-hour multi-pass method, using food habit and 24-hour recall questionnaires. Analysis of these diet data was undertaken using Dietplan-6 (version 6.70.43, 2013; Forestfield Software, Horsham, UK). The specific analysis methods are described in detail elsewhere [7]. We also objectively assessed diet using discriminative metabolite signatures, using an approach described in detail elsewhere [20]. Briefly, each participant's serum metabolite profile was obtained using a targeted metabolomics assay (Absolute*IDQ*$^{TM}$ p180 Kit, BIOCRATES Life Sciences AG, Innsbruck, Austria), which simultaneously quantifies 188 metabolites. In a previously published diet intervention study [20], serum samples had been collected in 19 participants who had undergone a metabolic ward-based supervised diet intervention. We assayed these blood samples using the BIOCRATES Absolute*IDQ*$^{TM}$ p180 Kit and derived diet-discriminatory metabolomic signatures, using previously described methods [20]. These data were then used to predict the dietary characteristics of the IMI DIRECT Study participants.

*Physical activity assessment:* Objective measures of physical activity (PA) were derived from triaxial accelerometers (ActiGraph GT3X+/GT3X+w/GT3X+bt, ActiGraph Co, Pensacola, USA) as described previously [7]. Raw data files (.gt3x) were converted to comma separated value (.csv) format storing rawest possible accelerations for each axis at a resolution of 30Hz using ActiLife 6 (version 6.11.5, ActiGraph Co, Pensacola, USA). All inferred measures of PA were calculated using PAMPRO (version uploaded 2015-10-21, MRC Epidemiology unit, Cambridge, UK), custom open source software available under public license (https://github.com/Thomite/pampro). Data from each axis of acceleration was auto-calibrated to local gravity. Non-wear was inferred as a vector magnitude standard deviation of less than 4mGs for a consecutive period greater than 60-min. All measures presented here have been adjusted for diurnal rhythm to account for bias from non-wear removal. However, due to the wear method (non-dominant wrist fastened using the manufacturer's non-removable hospital band), intermittent non-wear time was rare. The main PA estimates presented here are high-pass-filtered vector magnitude (hpfVM), which infers intensity of participants' movement in any direction at any given time (here, averaged during wear period). Time spent in established physical activity intensities by physical activity energy expenditure was estimated using calculated hpfVM cutpoints (Sedentary: (<48 mGs hpfVM, Light: 48-154 mGs hpfVM, Moderate: 154-389 mGs hpfVM, Vigorous: >389 mGs hpfVM). The methods used to infer these measures have been validated and described in detail elsewhere [21].

*DNA Extraction and Genotyping:* DNA extraction was carried out using Maxwell 16 Blood DNA purification kits and a Maxwell 16 semi-automated nucleic acid purification system (Promega). Genotyping was conducted using the Illumina HumanCore array (HCE24 v1.0) and genotypes were called using Illumina's GenCall algorithm. Samples were excluded for any of the following reasons: call rate <97%; low or excess mean heterozygosity; gender

discordance; and monozygosity. Genotyping quality control was then performed to provide high-quality genotype data for downstream analyses using the following criteria: call rate <99%; deviation from Hardy-Weinberg equilibrium (exact p<0.001); variants not mapped to human genome build GRCh37; and variants with duplicate chromosome positions (a total of 30,318 markers were excluded). A total of 3,032 samples and 517,958 markers across the two studies passed quality control procedures. We took autosomal variants with MAF>1% that passed quality control and constructed axes of genetic variation using principal components analysis implemented in the GCTA software to identify ethnic outliers defined as non-European ancestry using the 1000 Genomes samples as reference. We identified six individuals as ethnic outliers.

*Additional measures (not presented here):* Subsequent biomarker discovery analyses will employ additional measures (including 'omic' measures), which are outwith the scope of this cohort description. Additional measures that are not described here include transcriptomics (RNA sequencing from fasting whole blood), microbiomics (DNA isolation and deep sequencing in faecal samples), proteomics (targeted array in fasting plasma) and metabolomics (targeted and untargeted assays in fasting plasma). GAD/IA-2 assessments from fasting serum samples were also undertaken. Data from the Recent Physical Activity Questionnaire (RPAQ) and sleep diaries were also collected in sub-cohorts.

*Statistical power of study:* A detailed section on sample size and power for the study is available in the previously published rationale and design paper [7]. Briefly, statistical power will vary depending on a number of factors specific to the analysis to be carried out such as biomarker effect sizes, variance/frequency of outcome and biomarker, statistical modeling methods employed, number of tests (multiple testing adjustment), and of course sample size available for the relevant variables included in the model. The dataset will therefore be well

powered for some analyses while it may be underpowered for other analyses and will thus be covered in detail for the specific scenarios in subsequent analyses.

*Statistical methods for descriptive data:* Based on ADA 2011 diagnostic criteria, impaired fasting plasma glucose is 5.6-6.9mmol/l (100-125mg/dl), impaired glucose tolerance is 2-hr 75g OGTT plasma glucose 7.8-11mmol/l (140-199 mg/dl) and prediabetes $HbA_{1c}$ is 40-48mmol/mol (5.7-6.4%)[2]. Accordingly below and above these cut-offs was considered 'normal' and 'diabetic' ranges, respectively. Cohort 1 was stratified into two categories of glycaemia: normal glucose regulation (NGR) and impaired glucose regulation (IGR). NGR was defined as having $HbA_{1c}$, fasting glucose and 2-hr glucose values within the normal ranges for each measure. IGR was defined as having impaired values in at least one of $HbA_{1c}$, fasting glucose or 2-hr glucose. Cohort 2 was stratified into treatment categories: lifestyle advice only (LS) or metformin plus lifestyle advice (LS+MET). Descriptive statistics are presented as mean±SD. Pairwise Pearson correlations were calculated for all key variables described here. For these analyses, continuous variables were first inverse normal transformed and then adjusted for age, sex, and study centre by two-step residual regression. We present the same type of data for anthropometric and glycaemic variables for the main follow-up visits, as well as the differences for these variables between the baseline and the final follow-up visit measures (*follow-upΔ* = final follow-up value – baseline value). We also calculated pairwise Pearson correlations for the follow-upΔ values; for these analyses, continuous variables were first inverse normal transformed and then adjusted for age, sex, study centre and days since baseline visit by two-step residual regression. All statistics were computed using *R* software version 3.4.0 [22]. The IMI DIRECT data release version used for the analyses in this article was *direct_03-29-2019*.

*Glycaemic trait modeling:* Glycaemic traits were derived from frequently-sampled (0, 15, 30, 45, 60, 90, 120 min) 75g oral glucose tolerance tests (fsOGTT) or mixed-meal tolerance tests

(MMTT; 0, 30, 60, 90, 120 min) for Cohort 1 and Cohort 2, respectively. Analyses used a mathematical model that describes the relationship between insulin secretion and glucose concentration [23, 24]. The model expresses insulin secretion as the sum of two components: the first component represents the dependence of insulin secretion on absolute glucose concentration at any time during the fsOGTT/MMTT, through a dose-response function. Characteristic parameters of the dose-response relationship are the mean slope over the observed glucose range, denoted as *glucose sensitivity*. The dose-response relationship is modulated by a potentiation factor, which accounts for the fact that during acute stimulation, insulin secretion is higher on the descending phase of hyperglycaemia than at the same glucose concentration on the ascending phase. In participants with normal glucose regulation and insulin secretion, the potentiation factor typically increases from baseline to the end of a 2-hr OGTT [25]. To quantify this excursion, the ratio between the 2-hr and the baseline values were calculated. This ratio is denoted as *potentiation ratio* and reflects late insulin release. The second insulin secretion component represents the dependence of insulin secretion on the rate of change of glucose concentration. This component is termed *derivative component*, and is determined by a single parameter, denoted as *rate sensitivity*. Rate sensitivity reflects early insulin release [25].

The model parameters were estimated from glucose and C-peptide concentrations by regularized least-squares, as previously described [23]. Regularization involves the choice of smoothing factors, which were selected to obtain glucose and C-peptide model residuals with standard deviations close to expected measurement error (~1% for glucose and ~4% for C-peptide). Insulin secretion rates were calculated from the model every 5 minutes. The integral of insulin secretion during the fsOGTT was denoted as total insulin output.

The validity of the fsOGTT and MMTT for the assessment of insulin sensitivity has been shown in the original publications presenting the indices[26-28]. In the studies, the

OGTT/MMTT indices are compared with the euglycemic glucose clamp. The validity of the beta-cell function model is supported by numerous studies[25]. For beta-cell function, it is not possible to validate an OGTT/MMTT method against the classical tests with intravenous glucose infusion, due to the presence of the incretin effect. However, it has been shown that the estimated beta-cell dose-response is consistent with the graded glucose infusion test across the spectrum of glucose tolerance[29, 30].

**Results:**

Cohort 1 – Prediabetes Risk

Of 2235 enrolled participants in Cohort 1, 2127 passed all inclusion, exclusion, and quality control criteria. Of these, 1419 (67%) had IGR according to at least one ADA category for $HbA_{1c}$, fasting glucose or 2-hr glucose [2], and were thus within the target 'prediabetes' range. A total of 693 participants (33% of Cohort 1) had NGR for all three glycaemic measures. Participants with prevalent type 2 diabetes (n=105) or who withdrew from the study (n=3) were excluded from further analyses.

The number of participants enrolled into Cohort 1 varied between centres, with the Finnish sub-cohort being the largest with 58% (n=1240) of the total Cohort 1 baseline sample. The other centres in the Netherlands, Denmark and Sweden enrolled 22% (n=473), 13% (n=275), and 7% (n=139) of the total cohort, respectively.

The ratio of men to women varied in each sub-cohort, with all participants at the Finnish centre being male, and 43%, 45% and 29% being male in the sub-cohorts from Netherlands, Denmark and Sweden, respectively.

Detailed participant characteristics for baseline variables for Cohort 1 are shown in Table 1 (and stratified by glycaemic category in online ESM Table 1). Figure 2 shows the pairwise

correlations between a selection of key phenotypic variables at baseline. Participant characteristics at the follow-up visits and the difference (Δ) between baseline and final follow-up for Cohort 1 are shown in Table 2. The pairwise correlations between the baseline-final follow-up difference for anthropometric and glucose control variables are shown in Figure 3.

Briefly, at baseline, participants had a mean(SD) age of 62(6.2) years, BMI of 27.9(4.0) $kg/m^2$, $HbA_{1c}$ of 37(2.9) mmol/mol [5.5(0.27)%], fasting glucose of 5.7(0.6) mmol/l, 2-hr glucose of 5.9(1.6) mmol/l, fasting insulin of 10.9(7.6) pmol/l, glucose sensitivity of 113(55) pmol $min^{-1}m^{-2}mM^{-1}$, and insulin sensitivity (2-hr OGIS) of 381(59) ml $min^{-1}m^{-2}$. Participants had a 0 month – 48 month mean difference in fasting plasma glucose levels of 0.3(0.2) mmol/l.

Cohort 2 - New-onset diabetes

Of 830 patients enrolled to attend the screening visit in Cohort 2, 789 passed all inclusion, exclusion, and quality control criteria. Of these, 271 were treated with LS+MET and 508 were LS-only treated. Participants who withdrew consent, were receiving any other oral hypoglycaemic agent, or who reported ever receiving insulin treatment were been excluded (n=41).

Of those enrolled into Cohort 2, the UK (Dundee, Exeter, Newcastle), Dutch (Amsterdam), Swedish (Lund), and Danish (Copenhagen) study centres enrolled 21% (n=167), 18% (n=141), 21% (n=167), 20% (n=158), 12% (n=96) and 7% (n=52) of the total cohort, respectively; 52% to 63% of these sub-cohorts was male.

Detailed participant characteristics at baseline for key variables for Cohort 2 are shown in Table 1 (and stratified by treatment category in online ESM Table 1). Figure 2 shows the

pairwise correlation matrix for key variables at baseline adjusted for age, sex and study centre. Participant characteristics for follow-up visits and the difference between the baseline and final follow-up visit (Δ) for Cohort 2 is shown in Table 2. A pairwise correlation matrix for the anthropometric and glycaemic Δ variables are shown in online Figure 3.

Briefly, at baseline, participants had a mean(SD) age of 62(8.1) years, BMI of 30.5(5.0) $kg/m^2$, $HbA_{1c}$ of 46.5(5.8) mmol/mol [6.4(0.53)%], fasting glucose of 7.2(1.4) mmol/l, 2-hr glucose of 8.6(2.8) mmol/l, fasting insulin of 107(71) pmol/l, glucose sensitivity of 83(55) pmol $min^{-1}m^{-2}mM^{-1}$, and insulin sensitivity (2-hr OGIS) of 298(69) ml $min^{-1}m^{-2}$. Participants had a 0 month – 18 month mean difference in fasting plasma glucose levels of 0.8(1.9) mmol/l.

Genetic population substructure

As some study centres enrolled participants into both cohorts, we elected to characterize the genetic population substructure across the cohorts by study centre (i.e., pooling both cohorts at a given centre where possible). Genetic substructures closely map to the geographic location of the populations [31], indicating ethnic homogeneity within regions from which the cohorts were recruited, whereas there is far greater heterogeneity between centres, the latter driven mainly by the inclusion of Finnish participants. This is illustrated in Figure 4 where red circles denoting Finnish participants form a distinct cluster (to the north east) compared with the population from the other cohorts.

**Discussion**

Here, we report the characteristics of the IMI DIRECT cohorts at baseline and 18-months follow-up for glycaemic deterioration and consider these results in the context of the implemented protocols and the plans outlined in the design and rationale paper published

previously [7]. The descriptive statistics, pairwise correlations and genetic substructures presented in this article are not intended for etiological inference; instead, the purpose is to provide context and details for subsequent IMI DIRECT papers, as well as to inform scientists outside the consortium who might in the future consider using the IMI DIRECT data in their research.

Major advances in technologies and methods over the past decade make high-resolution quantification of disease phenotypes and processes possible in large sample collections. Applying modern assays to historical biosamples is particularly useful when studying processes that take decades to unfold. However, biosamples often degrade with long-term storage and many older studies did not deploy the advanced phenotyping methods available today. Recognizing these limitations, we designed and initiated two state-of-the-art prospective cohort studies as part of the IMI DIRECT consortium. Designed for biomarker discovery in glycaemic deterioration and diabetes progression, the IMI DIRECT cohorts include conventional and cutting-edge phenotyping techniques and technologies that are repeated on multiple occasions during a follow-up period of up to 48-months (currently on-going). We note that the subsequent biomarker discovery analyses using these cohorts will combine the clinical phenotypic data described in this paper with omic measures such as transcriptomics (RNA sequencing from fasting whole blood), microbiomics (DNA isolation and deep sequencing in faecal samples), proteomics (targeted array in fasting plasma) and metabolomics (targeted and untargeted assays in fasting plasma).

The recruitment strategies for the two IMI DIRECT cohorts differed in that Cohort 1 focused on recruiting from an existing large sample-frame (N=24,682) derived from established prospective cohort studies, whereas Cohort 2 used clinical registries to identify eligible participants. The strategy for recruiting participants from existing prospective studies for Cohort 1 facilitated access to data that was used to predict risk of rapid glycaemic

deterioration. However, despite the relatively large sampling-frame, it was necessary to enrol lower-risk participants in order to achieve the target sample size; in doing so, we recognized that this would likely reduce the overall rate of glycaemic deterioration in the cohort, although the generalizability of the study's findings will be greater. In Cohort 2, we fell slightly short of the target sample size of 1000 participants (n=789 with complete and high-quality data eventually enrolled), reflecting the difficulties in engaging some general practices, which was necessary to access diabetes registries in some regions.

We stratified Cohorts 1 and 2 by broad glycaemic category (overtly normoglycaemic or impaired glycaemic regulation in any ADA category) or treatment category (lifestyle only vs lifestyle + metformin), respectively, to reflect the basic stages of progression at baseline for descriptive purposes. Subsequent analyses will seek to further stratify by more disease pathogenic specific methods.

The two IMI DIRECT cohorts are not identical. However, they share many methodology parallels that permit complimentary analyses to be performed, such as determining whether biomarkers for glycaemic deterioration are conditional on disease state. However, several key differences in the protocols (e.g. fsOGTT vs MMTT) should be considered when interpreting results. We also note a discrepancy in the difference of missing of accelerometry (physical activity) data between Cohort 1 (19%) and Cohort 2 (8.5%); which we were unable to definitively explain. Partitioning change from error is very challenging when variables are assessed at only two time-points owing to regression to the mean. Not withstanding this, we note a modest 0 month – 48 month mean difference in fasting plasma glucose levels 0.3(0.5) mmol/l and 0.8(1.9) mmol/l for Cohort 1 and Cohort 2, respectively (Table 2) which likely reflect the relatively brief between visit interval. Furthermore we note that the 0.5 mmol/l and 1.9 mmol/l standard deviations in these differences, for Cohort 1 and Cohort 2 respectively, suggest the potential heterogeneity in changes in glycaemic control within each

cohort. With this in mind, the IMI-DIRECT cohorts are being followed further with record-linkage through 2026. We also note a discrepancy in the difference of missing of accelerometry (physical activity) data between Cohort 1 (19%) and Cohort 2 (8.5%). It should also be noted that the IMI DIRECT cohorts are predominantly of European ancestry therefore results from subsequent analyses on these cohorts will need to be replicated in other cohorts of relevant ancestry before generalising findings to other ethnicities. Finally, we note that the results presented here reflect the data available at publication.

## Conclusion

The study described here is being used to unravel the heterogeneous nature of glycaemic deterioration at risk of- and in diabetes, and to discover biomarkers that might prove useful for patient stratification and therapeutic optimization. As more prospective data are accrued, the IMI DIRECT cohorts will grow in value. In the long-term, the IMI DIRECT consortium intends to make these data available to other researchers through a managed-access repository.

**Funding:**

This work was supported by the Innovative Medicines Initiative Joint Undertaking under grant agreement n°115317 (DIRECT), resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and EFPIA companies' in kind contribution. R.W.K. was funded by a STAR Award Novo Nordisk co-financed PhD fellowship. The work undertaken by P.W.F was supported in part by ERC-2015-CoG_NASCENT_681742. E.R.P. holds a Wellcome Trust Investigator award (Grant reference 102820/Z/13/Z). Contributions to this work by S.Bru. were co-financed by the Novo Nordisk Foundation (grants NNF17OC0027594 and NNF14CC0001).

**Acknowledgements:**

We thank all the participants and study centre staff in IMI DIRECT for their contribution to the study. Some of the data were presented as an abstract at the Precision Medicine in Type 2 Diabetes and Cardiovascular Disease meeting in 2016.

**Author Contributions:**

R.W.K and P.W.F wrote the manuscript. P.W.F, M.W. and E.P. conceptualized and designed the studies with contributions from R.W.K, I.M.F, H.J.A.T., J.A., J.B., S.Bru., E.T.D., P.F., G.F., R.G., A.Ha., B.J., T.J.M., O.P., J.M.S., I.P., A.Mar., M.I.McM., and H.R. R.W.K. carried out the statistical analyses presented here. R.W.K, A.Ku., A.V., M.S., K.H.A., S.Bra., C.A.B., A.Y.D., F.DeM., C.J.G., T.K., A.M., M.H.P., S.P.R., E.L.T., A.T., T.W., J.A., J.B., S.Bru., E.T.D., P.F., G.F., R.G., M.L., T.J.M., O.P., J.M.S., A.M., M.I.McM., E.P., and P.W.F. contributed to sample assaying, data analysis/processing and/or data quality control procedures. R.W.K, I.M.F, A.He., G.N.G., T.H.H., M.H., A.Ko., F.R., M.S., T.K., M.R., H.J.A.T., H.V., J.D., T.H., A.Ha., M.L., O.P., M.W., E.P., and P.W.F. contributed to data

collection at study centres. P.W.F. coordinated the overall planning and implementation of the IMI DIRECT epidemiological studies. H.R. and E.P. were the overall coordinators of the IMI DIRECT consortium. All authors contributed to the revision of the manuscript. All authors critiqued and approved the final version of the manuscript. All authors accept responsibility for all aspects of the work insofar as ensuring that questions related to the accuracy or integrity of any part of the article are appropriately investigated and resolved. P.W.F. and E.P. are the guarantors of this work and, as such, had full access to the data described here and take responsibility for the integrity of the data and its analysis.

**Conflicts of interest:**

None

**References**

[1] Cho NH, Shaw JE, Karuranga S, et al. (2018) IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045. Diabetes Res Clin Pract 138: 271-281

[2] American Diabetes A (2011) Standards of medical care in diabetes--2011. Diabetes care 34 Suppl 1: S11-61

[3] Weyer C, Bogardus C, Mott DM, Pratley RE (1999) The natural history of insulin secretory dysfunction and insulin resistance in the pathogenesis of type 2 diabetes mellitus. The Journal of clinical investigation 104: 787-794

[4] Ringborg A, Lindgren P, Yin DD, Martinell M, Stalhammar J (2010) Time to insulin treatment and factors associated with insulin prescription in Swedish patients with type 2 diabetes. Diabetes & metabolism 36: 198-203

[5] Rathmann W, Strassburger K, Heier M, et al. (2009) Incidence of Type 2 diabetes in the elderly German population and the effect of clinical and lifestyle risk factors: KORA S4/F4 cohort study. Diabetic medicine : a journal of the British Diabetic Association 26: 1212-1219

[6] Ahlqvist E, Storm P, Karajamaki A, et al. (2018) Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. Lancet Diabetes & Endocrinology 6: 361-369

[7] Koivula RW, Heggie A, Barnett A, et al. (2014) Discovery of biomarkers for glycaemic deterioration before and after the onset of type 2 diabetes: rationale and design of the epidemiological studies within the IMI DIRECT Consortium. Diabetologia 57: 1132-1142

[8] Stancakova A, Javorsky M, Kuulasmaa T, Haffner SM, Kuusisto J, Laakso M (2009) Changes in insulin sensitivity and insulin release in relation to glycemia and glucose tolerance in 6,414 Finnish men. Diabetes 58: 1212-1221

[9] Hills SA, Balkau B, Coppack SW, et al. (2004) The EGIR-RISC STUDY (The European group for the study of insulin resistance: relationship between insulin sensitivity and cardiovascular disease risk): I. Methodology and objectives. Diabetologia 47: 566-570

[10] Rutters F, Nijpels G, Elders P, et al. (2018) Cohort Profile: The Hoorn Studies. International journal of epidemiology 47: 396-396j

[11] Christensen AI, Ekholm O, Glumer C, et al. (2012) The Danish National Health Survey 2010. Study design and respondent characteristics. Scandinavian journal of public health 40: 391-397

[12] Thuesen BH, Cerqueira C, Aadahl M, et al. (2014) Cohort Profile: the Health2006 cohort, research centre for prevention and health. International journal of epidemiology 43: 568-575

[13] DanFunD: Dansk undersøgelse af funktionelle lidelser 2016. Available from https://www.regionh.dk/fcfs/befolkningsbaseret-epidemiologi/befolkningsunders%C3%B8gelser/Sider/DanFund.aspx, accessed 05/03/2018 2018

[14] Roager HM, Vogt JK, Kristensen M, et al. (2017) Whole grain-rich diet reduces body weight and systemic low-grade inflammation without inducing major changes of the gut microbiome: a randomised cross-over trial. Gut

[15] Berglund G, Elmstahl S, Janzon L, Larsson SA (1993) The Malmo Diet and Cancer Study. Design and feasibility. Journal of internal medicine 233: 45-51

[16] Rauh SP, Heymans MW, Koopman AD, et al. (2017) Predicting glycated hemoglobin levels in the non-diabetic general population: Development and validation of the DIRECT-DETECT prediction model - a DIRECT study. PloS one 12: e0171816

[17] Thomas EL, Fitzpatrick JA, Malik SJ, Taylor-Robinson SD, Bell JD (2013) Whole body fat: content and distribution. Progress in nuclear magnetic resonance spectroscopy 73: 56-80

[18] O'Regan DP, Callaghan MF, Wylezinska-Arridge M, et al. (2008) Liver fat content and T2*: simultaneous measurement by using breath-hold multiecho MR imaging at 3.0 T--feasibility. Radiology 247: 550-557

[19] St Pierre TG, Clark PR, Chua-Anusorn W (2005) Measurement and mapping of liver iron concentrations using magnetic resonance imaging. Annals of the New York Academy of Sciences 1054: 379-385

[20] Garcia-Perez I, Posma JM, Gibson R, et al. (2017) Objective assessment of dietary patterns by use of metabolic phenotyping: a randomised, controlled, crossover trial. The lancet Diabetes & endocrinology 5: 184-195

[21] Brage S, Westgate K, Wijndaele K, Godinho J, Griffin S, Wareham N (2013) Evaluation of a method for minimising diurnal information bias in objective sensor data. Int Conf Amb Mon Phys Act Mov

[22] R-Core-Team (2014) R: A language and environment for statistical computing. . In. R Foundation for Statistical Computing, Vienna, Austria

[23] Mari A, Tura A, Gastaldelli A, Ferrannini E (2002) Assessing insulin secretion by modeling in multiple-meal tests: role of potentiation. Diabetes 51 Suppl 1: S221-226

[24] Mari A, Schmitz O, Gastaldelli A, Oestergaard T, Nyholm B, Ferrannini E (2002) Meal and oral glucose tests for assessment of beta -cell function: modeling analysis in normal subjects. American journal of physiology Endocrinology and metabolism 283: E1159-1166

[25] Mari A, Ferrannini E (2008) Beta-cell function assessment from modelling of oral tests: an effective approach. Diabetes, obesity & metabolism 10 Suppl 4: 77-87

[26] Mari A, Pacini G, Murphy E, Ludvik B, Nolan JJ (2001) A model-based method for assessing insulin sensitivity from the oral glucose tolerance test. Diabetes care 24: 539-548

[27] Matsuda M, DeFronzo RA (1999) Insulin sensitivity indices obtained from oral glucose tolerance testing: comparison with the euglycemic insulin clamp. Diabetes care 22: 1462-1470

[28] Stumvoll M, Mitrakou A, Pimenta W, et al. (2000) Use of the oral glucose tolerance test to assess insulin release and insulin sensitivity. Diabetes care 23: 295-301

[29] Ferrannini E, Gastaldelli A, Miyazaki Y, Matsuda M, Mari A, DeFronzo RA (2005) beta-Cell function in subjects spanning the range from normal glucose tolerance to overt diabetes: a new analysis. The Journal of clinical endocrinology and metabolism 90: 493-500

[30] Byrne MM, Sturis J, Sobel RJ, Polonsky KS (1996) Elevated plasma glucose 2 h postchallenge predicts defects in beta-cell function. The American journal of physiology 270: E572-579

[31] Novembre J, Johnson T, Bryc K, et al. (2008) Genes mirror geography within Europe. Nature 456: 98-101

**Tables and Figures:**

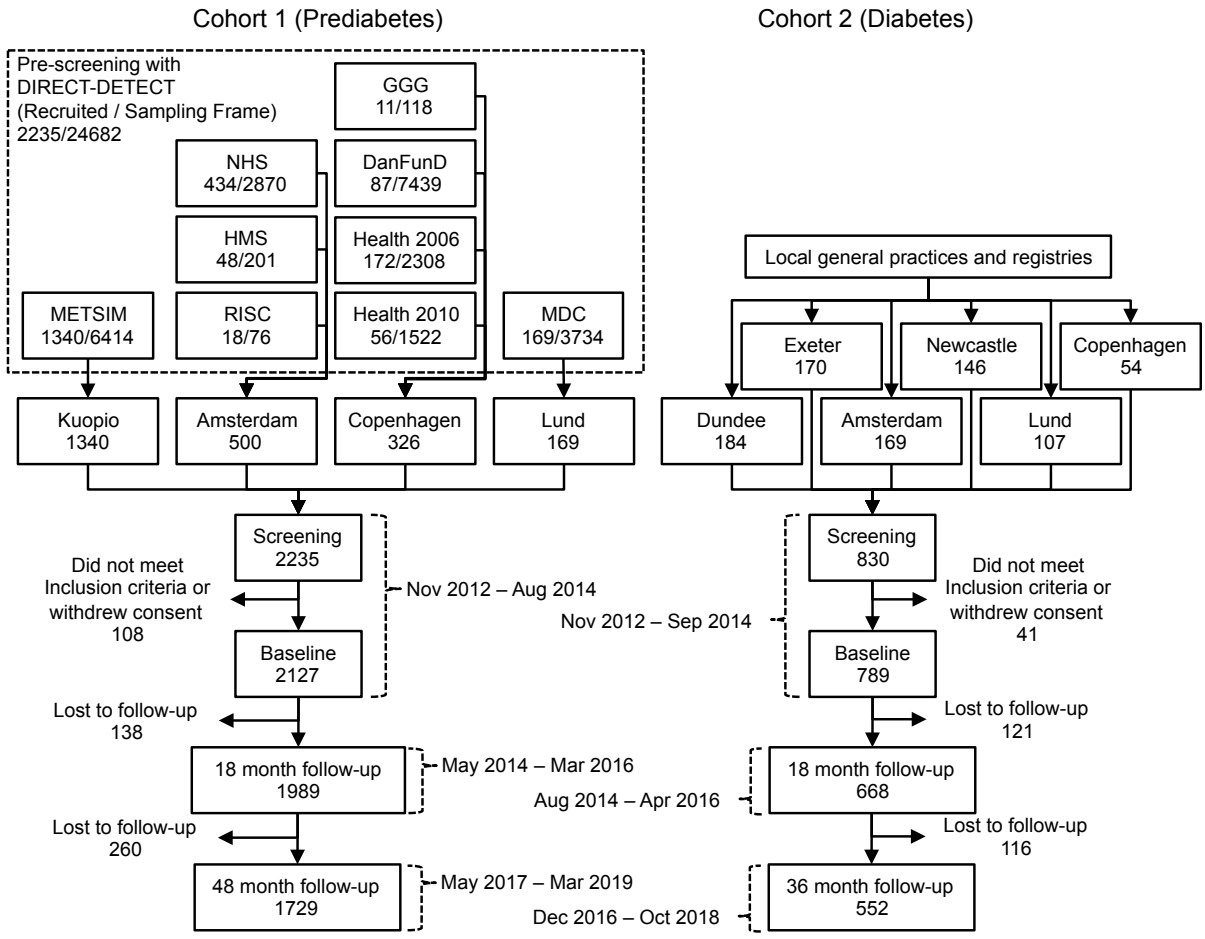


**Figure 1.** Participant flow of Cohorts 1 and 2. METSIM: Metabolic Syndrome in Men study, NHS: New Hoorn Study, HMS: Hoorn Meal Study, RISC: Relationship between Insulin Sensitivity and Cardiovascular disease cohort, GGG: Gut, Grain and Greens study, DanFunD: Danish Functional Disability study.

| | Cohort 1 (Prediabetes Risk) | | Cohort 2 (Diabetes) | |
|---|---|---|---|---|
| | Mean (SD) | N | Mean (SD) | N |
| **Male (%)** | 76% | 2127 | 58% | 789 |
| **Months since screening visit** | 6.4(4.8) | 2127 | 0.9(0.9) | 787 |
| **Age (yrs)** | 62(6.2) | 2127 | 62(8.1) | 787 |
| **Height (cm)** | 174(8) | 2127 | 171(9.8) | 787 |
| **Weight (kg)** | 85(13) | 2127 | 89(17) | 787 |
| **Waist circumference (cm)** | 99(11) | 2127 | 103(13) | 781 |
| **BMI ($kg \cdot m^{-2}$)** | 27.9(4.0) | 2127 | 30.5(5.0) | 787 |
| **Systolic blood pressure (mmHg)** | 131(15) | 2107 | 131(16) | 664 |
| **Diastolic blood pressure (mmHg)** | 81(9.0) | 2107 | 75(9.5) | 664 |
| **$HbA_{1c}$ ($mmol \cdot mol^{-1}$)** | 37(2.9) | 2113 | 46(5.8) | 784 |
| **$HbA_{1c}$ (%)** | 5.5(0.3) | 2113 | 6.4(0.5) | 784 |
| **Fasting glucose ($mmol \cdot L^{-1}$)** | 5.7(0.6) | 2126 | 7.2(1.4) | 787 |
| **Fasting insulin ($pmol \cdot L^{-1}$)** | 78.2(54.5) | 2124 | 106.6(70.9) | 787 |
| **Fasting HDL cholesterol ($mmol \cdot L^{-1}$)** | 1.3(0.4) | 2123 | 1.2(0.4) | 789 |
| **Fasting LDL cholesterol ($mmol \cdot L^{-1}$)** | 3.2(0.9) | 2123 | 2.3(1.0) | 781 |
| **Fasting triglycerides ($mmol \cdot L^{-1}$)** | 1.4(0.6) | 2123 | 1.5(0.9) | 789 |
| **Alanine aminotransferase ($IU \cdot L^{-1}$)** | 18(12) | 2120 | 26(14) | 789 |
| **Aspartate transaminase ($IU \cdot L^{-1}$)** | 27(10) | 2052 | 26(12) | 789 |
| **Total cholesterol ($mmol \cdot L^{-1}$)** | 5.1(1) | 2123 | 4.2(1.2) | 789 |
| **Fasting intact GLP-1 concentration ($pg \cdot mL^{-1}$)** | 0.41(0.59) | 2121 | 0.67(1.05) | 782 |
| **Fasting total GLP-1 concentration ($pg \cdot mL^{-1}$)** | 6.5(4.4) | 2120 | 9.4(9) | 780 |
| **Fasting glucagon ($pg \cdot mL^{-1}$)** | 98(41) | 2116 | 111(51) | 758 |
| **1hr intact proinsulin ($pg \cdot mL^{-1}$)** | 19(11.7) | 578 | 21(13.6) | 382 |
| **1hr GLP-1 increment ($pg \cdot mL^{-1}$)** | 9.3(12.1) | 2103 | 9.8(12.5) | 774 |
| **1hr glucagon increment ($pg \cdot mL^{-1}$)** | -10.7(38) | 2097 | -3.9(51) | 746 |
| **Mean 2hr glucose ($mmol \cdot L^{-1}$)** | 7.7(1.5) | 2126 | 9.3(2) | 779 |
| **Mean 2hr insulin ($pmol \cdot L^{-1}$)** | 383(260) | 2126 | 457(275) | 779 |
| **2hr glucose ($mmol \cdot L^{-1}$)** | 5.9(1.6) | 2127 | 8.6(2.8) | 786 |
| **2hr insulin ($pmol \cdot L^{-1}$)** | 48(48) | 2102 | 445(348) | 786 |
| **Fasting insulin secretion ($pmol \cdot min^{-1} \cdot m^{-2}$)** | 106(40) | 2126 | 137(48) | 779 |

| | | | | |
|---|---|---|---|---|
| **Integral of total insulin secretion ($nmol \cdot m^{-2}$)** | 52(18) | 2126 | 44(14) | 779 |
| **Glucose sensitivity ($pmol \cdot min^{-1} \cdot m^{-2} \cdot mmol^{-1} \cdot L^{-1}$)** | 113(55) | 2126 | 83(55) | 779 |
| **Rate sensitivity ($pmol \cdot m^{-2} \cdot mmol^{-1} \cdot L^{-1}$)** | 921(699) | 2126 | 1124(1082) | 779 |
| **Potentiation factor ratio (dimensionless)** | 1.7(0.6) | 2126 | 1.4(0.6) | 777 |
| **Insulin sensitivity 2hr OGIS ($ml \cdot min^{-1} \cdot m^{-2}$)** | 381(59) | 2118 | 298(69) | 775 |
| **Insulin sensitivity Stumvoll ($ml \cdot min^{-1} \cdot kg^{-1}$)** | 7.8(2.4) | 2099 | 5.5(2.7) | 775 |
| **Insulin sensitivity Matsuda (arbitrary)** | 5(3.1) | 2126 | 2.9(2.2) | 779 |
| **Intrabdominal Adipose Tissue (L)** | 5.5(2.4) | 956 | 5.7(2.2) | 374 |
| **Abdominal Subcutaneous Adipose Tissue (L)** | 6.1(2.6) | 953 | 8.1(3.8) | 374 |
| **Total Abdominal Adipose Tissue (L)** | 12(3.9) | 953 | 14(4.8) | 374 |
| **Liver Fat (%)** | 5(4.7) | 959 | 8.7(7.1) | 498 |
| **Pancreatic Fat (%)** | 13(8.9) | 929 | 11(7.3) | 446 |
| **Liver Iron content ($mg \cdot g^{-1}$)** | 1.3(0.26) | 958 | 1.4(0.31) | 498 |
| **Pancreatic Iron content ($mg \cdot g^{-1}$)** | 1.3(0.43) | 927 | 1.2(0.33) | 447 |
| **Average physical activity intensity - hpfVM (mGs)** | 37(10.2) | 1714 | 34(9.9) | 722 |
| **Sedentary (% of time)** | 82(4.2) | 1714 | 83(4.3) | 722 |
| **Light (% of time)** | 10.9(2.3) | 1714 | 10.4(2.3) | 722 |
| **Moderate (% of time)** | 5.3(1.5) | 1714 | 4.9(1.6) | 722 |
| **Vigorous (% of time)** | 1.5(0.7) | 1714 | 1.3(0.7) | 722 |
| **Total Energy ($kCal \cdot day^{-1}$)** | 1963(751) | 2064 | 1840(602) | 707 |
| **Carbohydrate ($g \cdot day^{-1}$)** | 223(96) | 2064 | 213(78) | 707 |
| **Fat intake ($g \cdot day^{-1}$)** | 79(39) | 2064 | 72(33) | 707 |
| **Protein ($g \cdot day^{-1}$)** | 99(44) | 2064 | 87(31) | 707 |
| **Sugar ($g \cdot day^{-1}$)** | 96(53) | 2064 | 85(43) | 707 |
| **Fibre ($g \cdot day^{-1}$)** | 20(9.8) | 2064 | 19(8.4) | 707 |
| **Saturated fat ($g \cdot day^{-1}$)** | 29(16) | 2064 | 26(14) | 707 |
| **Monunsaturated fat ($g \cdot day^{-1}$)** | 27(17) | 2064 | 24(13) | 707 |
| **Polyunsaturated fat ($g \cdot day^{-1}$)** | 13(8.2) | 2064 | 12(8) | 707 |

**Table 1.** Baseline clinical and phenotypic characteristics of Cohorts 1 and 2. Descriptive statistics shown are N, % or mean (standard deviation) as indicated. Values are untransformed and unadjusted. Data presented reflects the data available at the time of publication.

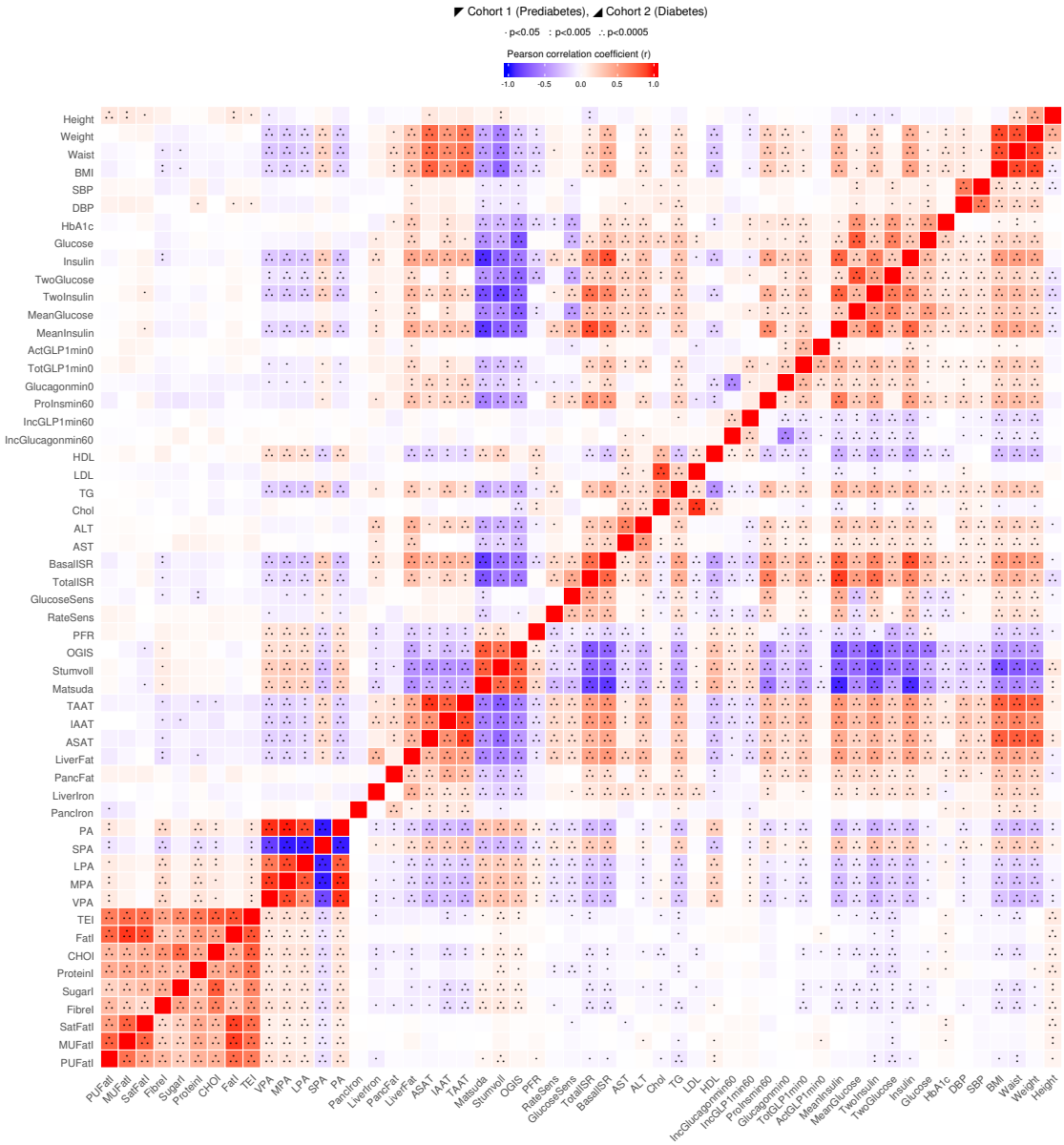


**Figure 2**. Pairwise correlation matrix. Fill color indicates Pearson correlation coefficient (r), where positive is denoted by red fill, inverse by blue fill, magnitude by intensity. Cohort 1 and 2 are separate, above and below diagonal, respectively. All continuous variables were normally transformed and adjusted for age, sex, and study center. SBP: Systolic blood pressure, DBP: Diastolic blood pressure, Glucose: Fasting glucose, Insulin: Fasting insulin, TwoGlucose: 2-hr glucose, TwoInsulin: 2-hr insulin, MeanGlucose: Mean 2-hr glucose, MeanInsulin: Mean 2-hr insulin, ActGLP1min0: Fasting intact GLP-1 concentration, TotGLP1min0: Fasting total GLP-1 concentration, Glucagonmin0: Fasting glucagon,

ProInsmin60: 1h intact proinsulin, IncGLP1min60: 1h GLP-1 increment, IncGlucagonmin60: 1h glucagon increment, HDL: Fasting HDL cholesterol, LDL: Fasting LDL cholesterol, TG: Fasting triglycerides, Chol: Total cholesterol, ALT: Alanine aminotransferase, AST: Aspartate transaminase, BasalISR: Fasting insulin secretion, TotalISR: Integral of total insulin secretion, GlucoseSens: Glucose sensitivity, RateSens: Rate sensitivity, PFR: Potentiation factor ratio, OGIS: Insulin sensitivity 2-h OGIS, Stumvoll: Insulin sensitivity Stumvoll, Matsuda: Insulin sensitivity Matsuda, IAAT: Intrabdominal Adipose Tissue, ASAT: Abdominal Subcutaneous Adipose Tissue, TAAT: Total Abdominal Adipose Tissue, LiverFat: Liver Fat, PancFat: Pancreatic Fat, LiverIron: Liver Iron content, PancIron: Pancreatic Iron content, PA: Average physical activity intensity - hpfVM, SPA: Sedentary, LPA: Light, MPA: Moderate, VPA: Vigorous, TEI: Total energy intake, FatI: Fat intake, CHOI: Carbohydrate intake, ProteinI: Protein intake, SugarI: Sugar intake, FibreI: Fibre intake, SatFatI: Saturated fat intake, MUFatI: Monunsaturated fat intake, PUFatI: Polyunsaturated fat intake.

| | Cohort 1 (Prediabetes) | | | | | Cohort 2 Diabetes) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 18 Month Mean (SD) | N | 48 Month Mean (SD) | N | Baseline to 48 Month Δ Mean (SD) | 18 Month Mean (SD) | N | 36 Month Mean (SD) | N | Baseline to 36 Month Δ Mean (SD) |
| **Male (%)** | 0.77 | 1989 | 0.79 | 1729 | | 0.6 | 668 | 0.61 | 552 | |
| **Months since previous visit** | 18.6(1.4) | 1989 | 30.8(1.3) | 1729 | | 18.2(0.6) | 668 | 18.1(1.0) | 552 | |
| **Age (yrs)** | 63.6(6.1) | 1989 | 65.9(6.1) | 1729 | | 63.8(7.7) | 668 | 65.4(7.7) | 552 | |
| **Height (cm)** | 174(7.9) | 1983 | 175(7.9) | 1729 | -0.39(1.01) | 171(9.8) | 662 | 171(9.9) | 551 | -0.5(1.29) |
| **Weight (kg)** | 85(13.5) | 1981 | 85.3(13.7) | 1729 | -0.16(4.6) | 89.6(16.9) | 664 | 89.3(16.6) | 548 | -0.13(5.5) |
| **Waist circumference (cm)** | 99(11) | 1981 | 100(11) | 1728 | 0.8(5.9) | 103(13) | 661 | 104(13) | 546 | 1.8(6.4) |
| **BMI ($kg \cdot m^{-2}$)** | 27.9(4) | 1980 | 28(4.1) | 1729 | 0.07(1.6) | 30.5(5) | 661 | 30.5(4.9) | 548 | 0.13(1.9) |
| **Systolic blood pressure (mmHg)** | 131(16) | 1983 | 132(16) | 1728 | 1.7(14) | 130(16) | 667 | 132(16) | 550 | 0.6(15) |
| **Diastolic blood pressure (mmHg)** | 80(8.8) | 1983 | 81(9.7) | 1728 | 0.1(8.2) | 74(9.3) | 665 | 74(9.6) | 548 | -1.2(8.9) |
| **HbA1c ($mmol \cdot mol^{-1}$)** | 37.9(3.1) | 1980 | 40.3(3.4) | 1668 | 3.2(2.3) | 48.5(8.9) | 662 | 48.1(9.9) | 545 | 1.7(8.8) |
| **HbA1c (%)** | 5.6(0.3) | 1980 | 5.8(0.3) | 1668 | 0.3(0.2) | 6.6(0.8) | 662 | 6.6 (0.9) | 545 | 0.2(0.8) |
| **Fasting glucose ($mmol \cdot L^{-1}$)** | 5.8(0.6) | 1977 | 6.0(0.6) | 1679 | 0.3(0.5) | 7.8(1.8) | 658 | 7.9(2) | 509 | 0.8(1.9) |
| **Fasting insulin ($pmol \cdot L^{-1}$)** | 84.7(61.7) | 1975 | 85.4(59.6) | 1713 | 8.6(42.3) | 118.8(78.6) | 655 | 119.3(75.4) | 496 | 12(62.9) |
| **2-hr glucose ($mmol \cdot L^{-1}$)** | 6.1(1.7) | 1975 | 6.5(2) | 1673 | 0.6(1.7) | 9.5(3.3) | 659 | 9.9(3.4) | 505 | 1.3(3.2) |
| **2-hr insulin ($pmol \cdot L^{-1}$)** | 55.3(53.7) | 1950 | 59(57.6) | 1687 | 11.2(43.5) | 449.7(313.4) | 648 | 445.7(325.2) | 488 | 10.4(289.8) |
| **Mean 2-hr glucose ($mmol \cdot L^{-1}$)** | 8(2) | 1974 | 8(2) | 1054 | 0.4(1.3) | 10(3) | 650 | 10(3) | 494 | 1.1(2.3) |
| **Mean 2-hr insulin ($pmol \cdot L^{-1}$)** | 425(280) | 1974 | 437(287) | 1054 | 37.2(190) | 476(278) | 650 | 466(283) | 494 | 6.7(207) |
| **Fasting insulin secretion ($pmol \cdot min^{-1} \cdot m^{-2}$)** | 103(41) | 1974 | 113(44) | 1054 | 5.7(29) | 145(52) | 650 | 148(54) | 494 | 10.7(41) |
| **Glucose sensitivity ($pmol \cdot min^{-1} \cdot m^{-2} \cdot mmol^{-1} \cdot L^{-1}$)** | 109(55) | 1974 | 114(55) | 1054 | 2.3(54.3) | 81(57) | 650 | 77(62) | 494 | -6.3(53.1) |
| **Rate sensitivity ($pmol \cdot m^{-2} \cdot mmol^{-1} \cdot L^{-1}$)** | 833(673) | 1974 | 859(627) | 1054 | -65(627) | 1176(1290) | 650 | 1011(871) | 494 | -56(1181) |
| **Potentiation factor ratio (dimensionless)** | 2(1) | 1973 | 2(1) | 1052 | 0.1(0.8) | 1(1) | 650 | 1(1) | 491 | 0(0.7) |
| **Integral of total insulin secretion ($nmol \cdot m^{-2}$)** | 50(19) | 1974 | 54(18) | 1054 | 1.5(12.6) | 45(14) | 650 | 45(14) | 494 | 1(11.4) |
| **Insulin sensitivity 2-h OGIS ($ml \cdot min^{-1} \cdot m^{-2}$)** | 369(60.8) | 1963 | 353(58.6) | 1046 | -21(46.1) | 277(53.9) | 641 | 276(53.9) | 476 | -23.2(71) |
| **Insulin sensitivity Stumvoll ($ml \cdot min^{-1} \cdot kg^{-1}$)** | 7.6(2.6) | 1941 | 7.4(2.7) | 1041 | -0.4(1.8) | 5.3(2.7) | 640 | 5.2(2.7) | 474 | -0.4(2.1) |
| **Insulin sensitivity Matsuda (arbitrary)** | 4.7(3.1) | 1974 | 4.4(2.8) | 1054 | -0.6(2.2) | 2.4(1.7) | 649 | 2.4(1.6) | 491 | -0.5(1.9) |

**Table 2.** Follow-up and Follow-up **Δ** in clinical and phenotypic characteristics of Cohorts 1 and 2. Descriptive statistics shown are N, % or mean (standard deviation) as indicated. Values are untransformed and unadjusted. Follow-up **Δ** is mean (standard deviation) of difference between characteristic value between follow-up assessment and baseline visits.

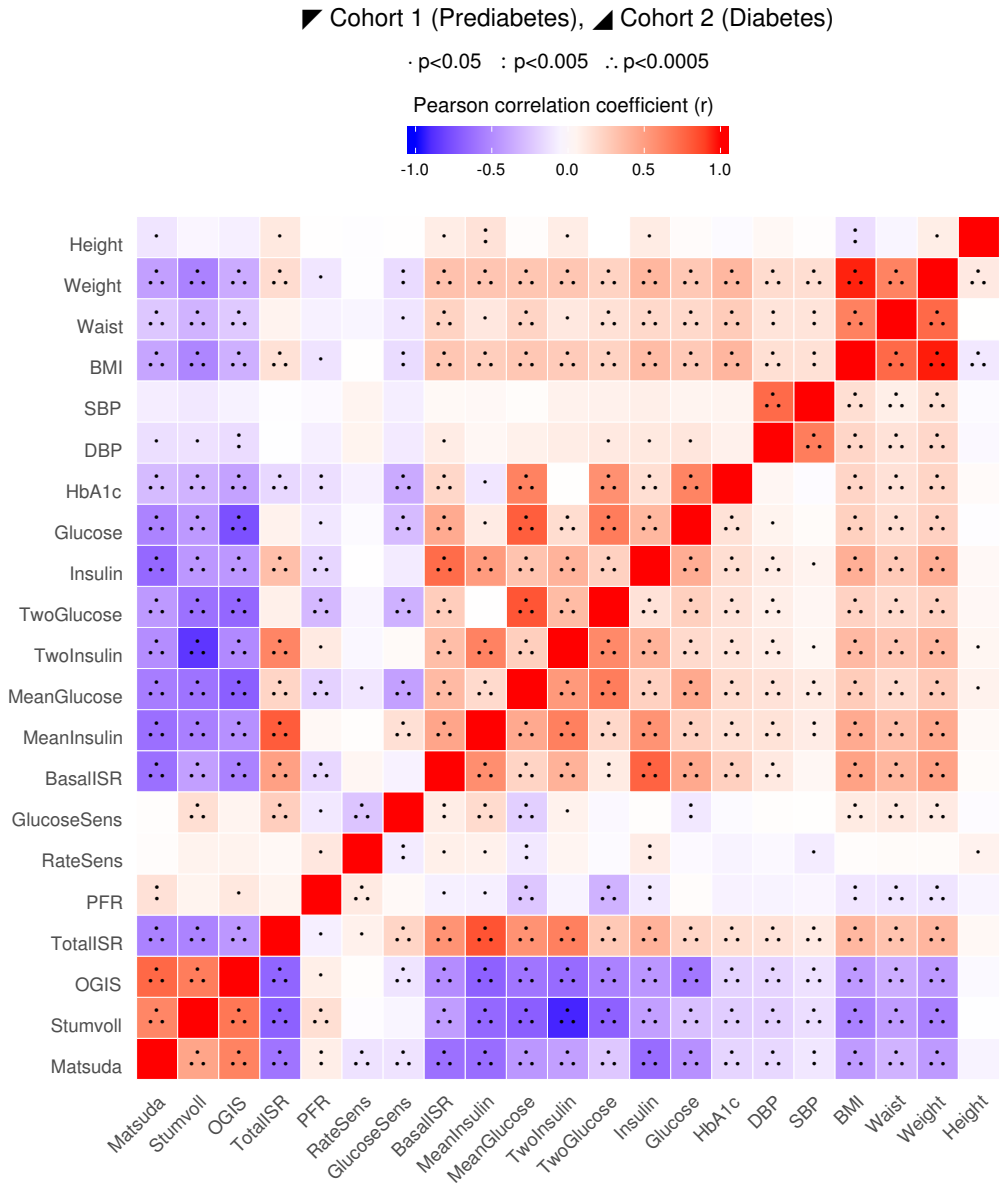


**Figure 3**. Pairwise correlation matrix of Follow-up **Δ.** (difference between characteristic value between 48 follow-up assessment and baseline visits). Fill color indicates Pearson correlation coefficient (r), where positive is denoted by red fill, inverse by blue fill, magnitude by intensity. Cohort 1 and 2 are separate, above and below diagonal, respectively. All continuous variables were normally transformed and adjusted for age, sex, and study center. SBP: Systolic blood pressure, DBP: Diastolic blood pressure, Glucose: Fasting glucose, Insulin: Fasting insulin, TwoGlucose: 2-hr glucose, TwoInsulin: 2-hr insulin, MeanGlucose: Mean 2-hr glucose, MeanInsulin: Mean 2-hr insulin, BasalISR: Fasting insulin

secretion, TotalISR: Integral of total insulin secretion, GlucoseSens: Glucose sensitivity, RateSens: Rate sensitivity, PFR: Potentiation factor ratio, OGIS: Insulin sensitivity 2-h OGIS, Stumvoll: Insulin sensitivity Stumvoll, Matsuda: Insulin sensitivity Matsuda.

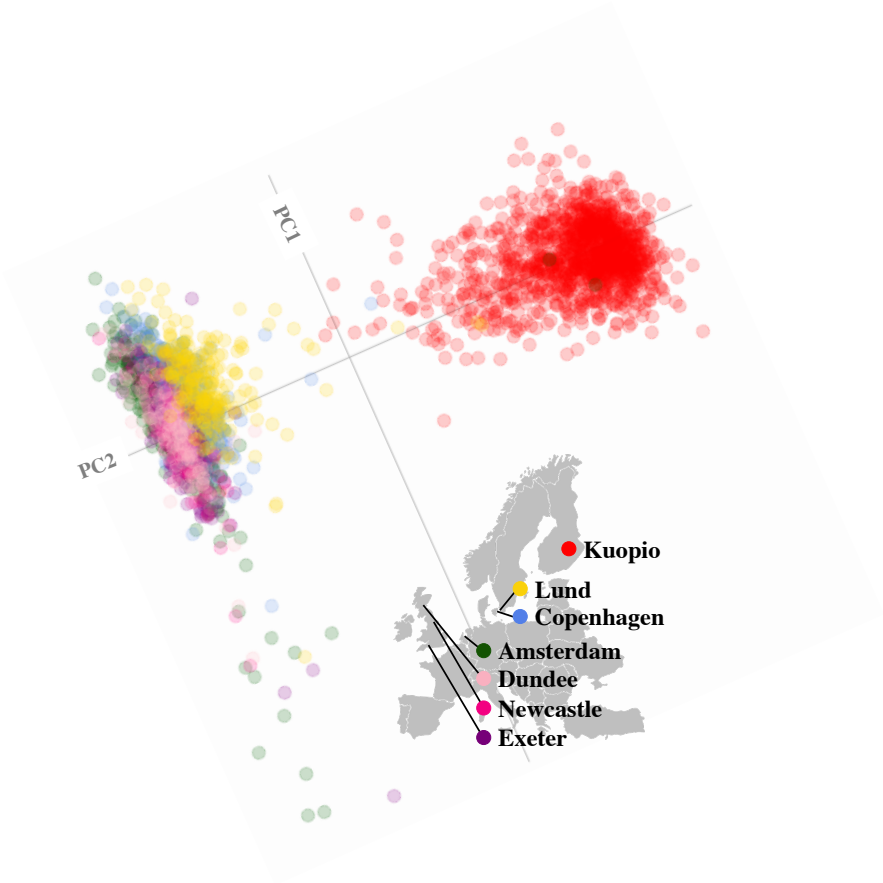


**Figure 4**. Genetic population structure within WP2 baseline study samples. A statistical summary of genetic data from Cohorts 1 and 2 (combined) based on principal component axis one (PC1) and axis two (PC2). Points are colored as per the recruitment centers. Red, Kuopio; Yellow, Lund; Blue, Copenhagen; Green, Amsterdam; Pink, Newcastle; Salmon, Dundee; Purple, Exeter.

## ESM Table 1.

| | | Cohort 1 (Prediabetes) | | | Cohort 2 (Diabetes) | | |
|---|---|---|---|---|---|---|---|
| | | NGR | IGR | NGR/IGR | LS | LS+MET | NGR/IGR |
| | | Mean (SD) | Mean (SD) | N | Mean (SD) | Mean (SD) | N |
| Anthropometrics | Male (%) | 75% | 76% | 693, 1419 | 58% | 56% | 508, 271 |
| | Months since previous visit | 6.8(4.9) | 6.2(4.8) | 693, 1419 | 0.8(1) | 0.9(0.8) | 507, 271 |
| | Age (yrs) | 62(6.2) | 62(6.2) | 693, 1419 | 62(7.9) | 61(8.3) | 507, 271 |
| | Height (cm) | 174(7.9) | 174(8.1) | 693, 1419 | 171(9.5) | 171(10.2) | 507, 271 |
| | Weight (kg) | 83(12) | 86(14) | 693, 1419 | 89(17) | 89(17) | 507, 271 |
| | Waist circumference (cm) | 97(10) | 101(11) | 693, 1419 | 103(14) | 104(13) | 502, 270 |
| | BMI ($kg \cdot m^{-2}$) | 27.2(3.7) | 28.3(4.1) | 693, 1419 | 30.7(5.1) | 30.3(4.8) | 507, 271 |
| | Systolic blood pressure (mmHg) | 129(15) | 132(15) | 685, 1407 | 132(16) | 129(14) | 450, 211 |
| | Diastolic blood pressure (mmHg) | 79(8.7) | 81(9) | 685, 1407 | 75(9.6) | 75(9.3) | 450, 211 |
| Glycaemic Control | HbA1c (mmol·mol-1) | 35(2.2) | 38(2.9) | 693, 1419 | 46(5.7) | 47(5.8) | 508, 266 |
| | HbA1c (%) | 5.4(0.2) | 5.6(0.3) | 693, 1419 | 6.4(0.5) | 6.5(0.5) | 508, 266 |
| | Fasting glucose (mmol·L-1) | 5.2(0.4) | 5.9(0.5) | 693, 1419 | 7(1.4) | 7.5(1.5) | 506, 271 |
| | Fasting insulin (pmol·L-1) | 60.3(36.3) | 86.8(60.3) | 693, 1417 | 105.8(70.6) | 108.1(72) | 506, 271 |
| | Fasting HDL cholesterol (mmol·L-1) | 1.4(0.3) | 1.3(0.4) | 690, 1418 | 1.2(0.4) | 1.2(0.4) | 508, 271 |
| | Fasting LDL cholesterol (mmol·L-1) | 3.2(0.9) | 3.2(0.9) | 690, 1418 | 2.3(1) | 2.3(0.9) | 506, 266 |
| | Fasting triglycerides (mmol·L-1) | 1.2(0.5) | 1.4(0.7) | 690, 1418 | 1.5(0.8) | 1.7(0.9) | 508, 271 |
| | Alanine aminotransferase (IU·L-1) | 16(10) | 19(13) | 690, 1415 | 25(12) | 28(16) | 508, 271 |
| | Aspartate transaminase (IU·L-1) | 26(9) | 28(11) | 669, 1369 | 25(10) | 27(14) | 508, 271 |
| | Total cholesterol (mmol·L-1) | 5.1(1) | 5.1(1) | 690, 1418 | 4.2(1.2) | 4.3(1.1) | 508, 271 |
| | Fasting intact GLP-1 concentration (pg·mL-1) | 0.37(0.45) | 0.42(0.65) | 691, 1415 | 0.61(1.04) | 0.77(1.06) | 505, 267 |
| | Fasting total GLP-1 concentration (pg·mL-1) | 6.1(4.1) | 6.7(4.5) | 690, 1415 | 9(7.6) | 10.1(11.2) | 502, 268 |
| | Fasting glucagon (pg·mL-1) | 96(33) | 100(44) | 690, 1411 | 112(56) | 109(38) | 488, 261 |
| | 1h intact proinsulin (pg·mL-1) | 15(7.4) | 20(12.7) | 145, 430 | 22(14) | 20(12.9) | 251, 128 |
| | 1h GLP-1 increment (pg·mL-1) | 9.4(11.4) | 9.3(12.4) | 687, 1401 | 9.9(14) | 9.4(8.9) | 497, 267 |
| | 1h glucagon increment (pg·mL-1) | -9.5(52) | -11.3(29) | 684, 1398 | -4.8(58) | -1.9(37) | 479, 259 |
| | Mean 2-hr glucose (mmol·L-1) | 6.8(1.1) | 8.2(1.4) | 692, 1419 | 8.9(1.9) | 10.1(2.1) | 503, 267 |
| | Mean 2-hr insulin (pmol·L-1) | 314(203) | 418(279) | 692, 1419 | 469(289) | 433(244) | 503, 267 |
| | 2-hr glucose (mmol·L-1) | 5.3(1.2) | 6.2(1.8) | 693, 1419 | 8.2(2.7) | 9.5(2.9) | 505, 271 |
| | 2-hr insulin (pmol·L-1) | 35(31) | 55(53) | 682, 1405 | 455(347) | 428(352) | 505, 271 |
| | Fasting insulin secretion (pmol·min-1·m-2) | 90(30) | 114(42) | 692, 1419 | 134(48) | 141(48) | 503, 267 |
| | Integral of total insulin secretion (nmol·m-2) | 47(15) | 55(18) | 692, 1419 | 44(15) | 44(13) | 503, 267 |
| | Glucose sensitivity (pmol·min-1·m-2·mmol-1·L-1) | 124(62) | 107(51) | 692, 1419 | 90(57) | 72(50) | 503, 267 |
| | Rate sensitivity (pmol·m-2·mmol-1·L-1) | 1003(768) | 880(658) | 692, 1419 | 1122(1119) | 1130(1018) | 503, 267 |
| | Potentiation factor ratio (dimensionless) | 1.8(0.6) | 1.7(0.6) | 692, 1419 | 1.5(0.7) | 1.3(0.4) | 501, 267 |
| | Insulin sensitivity 2-h OGIS (ml·min-1·m-2) | 417(49) | 362(56) | 689, 1414 | 303(74) | 288(60) | 500, 266 |
| | Insulin sensitivity Stumvoll (ml·min-1·kg-1) | 8.7(1.7) | 7.4(2.6) | 681, 1403 | 5.6(2.8) | 5.4(2.6) | 500, 266 |
| | Insulin sensitivity Matsuda (arbitrary) | 6.5(3.6) | 4.3(2.5) | 692, 1419 | 3(2.3) | 2.7(1.9) | 503, 267 |
| Magnetic Resonance Imaging | Intrabdominal Adipose Tissue (L) | 5.1(2.3) | 5.6(2.4) | 274, 674 | 6(2.2) | 5.2(2.1) | 247, 122 |
| | Abdominal Subcutaneous Adipose Tissue (L) | 5.5(2.3) | 6.3(2.7) | 273, 672 | 8.1(3.7) | 7.9(3.9) | 247, 122 |
| | Total Abdominal Adipose Tissue (L) | 11(3.6) | 12(4) | 273, 672 | 14(4.7) | 13(4.9) | 247, 122 |
| | Liver Fat (%) | 3.8(3.6) | 5.5(5) | 276, 675 | 8.2(6.6) | 9.9(8) | 340, 153 |
| | Pancreatic Fat (%) | 14(8.8) | 13(9) | 262, 660 | 11(6.7) | 11(8.2) | 307, 136 |
| | Liver Iron content (mg·g-1) | 1.3(0.31) | 1.3(0.24) | 276, 674 | 1.4(0.28) | 1.4(0.35) | 340, 153 |
| | Pancreatic Iron content (mg·g-1) | 1.3(0.63) | 1.3(0.31) | 262, 658 | 1.2(0.33) | 1.2(0.31) | 306, 138 |
| Physical Activity | Average physical activity intensity - hpfVM (mgs) | 38(10.5) | 37(10) | 561, 1140 | 35(10.1) | 33(9.4) | 465, 248 |
| | Sedentary (% of time) | 82(4.4) | 82(4.1) | 561, 1140 | 83(4.4) | 84(4.1) | 465, 248 |
| | Light (% of time) | 10.9(2.4) | 10.8(2.2) | 561, 1140 | 10.6(2.4) | 10.1(2.2) | 465, 248 |
| | Moderate (% of time) | 5.3(1.6) | 5.2(1.5) | 561, 1140 | 5(1.6) | 4.7(1.5) | 465, 248 |
| | Vigorous (% of time) | 1.6(0.8) | 1.5(0.7) | 561, 1140 | 1.4(0.7) | 1.2(0.6) | 465, 248 |
| Diet | Protein (g·day-1) | 1974(723) | 1962(765) | 670, 1380 | 1825(600) | 1865(609) | 474, 224 |
| | Carbohydrate (g·day-1) | 227(95) | 221(97) | 670, 1380 | 212(80) | 215(73) | 474, 224 |
| | Fat intake (g·day-1) | 79(38) | 79(39) | 670, 1380 | 71(32) | 74(35) | 474, 224 |
| | Total Energy (kCal·day-1) | 98(43) | 99(45) | 670, 1380 | 86(29) | 89(35) | 474, 224 |
| | Sugar (g·day-1) | 100(53) | 95(54) | 670, 1380 | 84(42) | 87(45) | 474, 224 |
| | Fibre (g·day-1) | 21(10.1) | 20(9.6) | 670, 1380 | 20(8.9) | 18(7.2) | 474, 224 |
| | Saturated fat (g·day-1) | 29(16) | 30(17) | 670, 1380 | 25(14) | 28(14) | 474, 224 |
| | Monunsaturated fat (g·day-1) | 28(18) | 27(16) | 670, 1380 | 24(12) | 25(13) | 474, 224 |
| | Polyunsaturated fat (g·day-1) | 13(8.6) | 13(8.1) | 670, 1380 | 12(7.6) | 12(8.7) | 474, 224 |

**ESM Table 1.** Baseline clinical and phenotypic characteristics of Cohorts 1 and 2 by glycaemic control and treatment status. IGR: Impaired glucose regulation, NGR: Normal glucose regulation, LS: Lifestyle Treatment, LS+MET: Lifestyle and metformin treatment. Descriptive statistics shown are N, % or mean (standard deviation) as indicated. Values are untransformed and unadjusted. Data presented reflects the data available at the time of publication.